Hydrology and Earth System Sciences, 3(2), *305–307* (1999) © EGS

Hydrology & Earth System Sciences

RESEARCH NOTE

Hydrograph Prediction—How much skill?

Max Beran

OB Research Services, The Croft, East Hagbourne, Didcot, Oxon OX11 9LS
e-mail for corresponding author: max@oldboot.demon.co.uk

Abstract

The matching of estimated to observed hydrograph shape is central to much hydrological analysis. This research note quantifies built-in biases that tend to inflate goodness of fit indicies, biases that arise from the similarity of geometry between observed and estimated hydrographs.

Introduction

How much genuine skill is displayed by hydrograph prediction techniques? This paper shows that the answer is, 'a lot less than is commonly supposed', a conclusion which is based on substituting a more realistic 'null hypothesis' for the ones used currently in testing the significance of hydrograph predictions.

A typical instance occurs with flood analysis by means of the unit hydrograph. The accuracy of the temporal aspects of the procedure is judged by applying the unit hydrograph to observed flood events, essentially by convoluting it with an observed effective precipitation hyetograph and comparing the predicted with the recorded immediate response hydrograph. Objective functions have been proposed for hydrograph simulation models based on explained variance and on likelihood (for example Nash, J.E. and Sutcliffe, 1970; Binley *et al.*, 1991; Jakeman, A.J. and Hornberger, G.M., 1993; Gupta, H.V. *et al.*, 1998). However, this past work has been directed mainly to parameter optimisation and model identification. The question of how much skill is represented by a given value of the objective function is not addressed directly.

To answer the question in the title, it is necessary to set a baseline or null-hypothesis with which a test statistic can be compared. An inappropriate null-hypothesis has led in the past to a tendency to exaggerate the significance of the hydrograph fit and hence to exaggerate the real level of modelling skill. Though expressed here in the context of unit hydrograph analysis, this research note's conclusions apply to other hydrological input:output models in which a transfer function is evaluated by measuring how closely it is able to replicate the shape (as opposed to the volume) of a recorded hydrograph.

Evaluating goodness of fit

Two measures of goodness of fit are considered—the Nash and Sutcliffe (1970) Coefficient of Efficiency, E and the product moment correlation coefficient, R. Both E and R are measures of explained variance and may be regarded as quite severe tests inasmuch as they operate on corresponding hydrograph ordinates and so make no allowance for small temporal shifts. Values of E or R in the region of 0.8 or 0.9 are often regarded as significant, meaning that the transfer function model is safe to apply.

$$E = 1 - \left\{ \Sigma(Q_{\mathrm{rec}} - Q_{\mathrm{est}})^2 \;/\; \Sigma(Q_{\mathrm{rec}} - Q_{\mathrm{mean}})^2 \right\} \tag{1}$$

where Q_{rec}, Q_{est} and Q_{mean} are the recorded, estimated and mean values of discharge and the summation is taken over all ordinates within the hydrograph range.

While at first sight such values do suggest a high level of skill, in the case of the coefficient E, this measure is not normalized so there is no simple relationship with R or R^2, the conventional statistical measure of explained variance. Thus $E = 0$ can occur simultaneously with high positive correlation between Q_{rec} and Q_{est}, see Fig. 1. Also, as can be seen by inspection of Eqn. 1, large negative values can build up. This may occur even with moderate positive correlation between the hydrograph ordinates when the estimated values depart from the recorded value more than the latter do from their own mean value. There is also an element of arbitrariness in the calculation as both measures can be inflated by the inclusion of lengths of lead-in and hydrograph tail within the summation range.

These difficulties are ones of implementation. The more important matter, with which this paper is primarily concerned, is of a conceptual nature, and lies with the

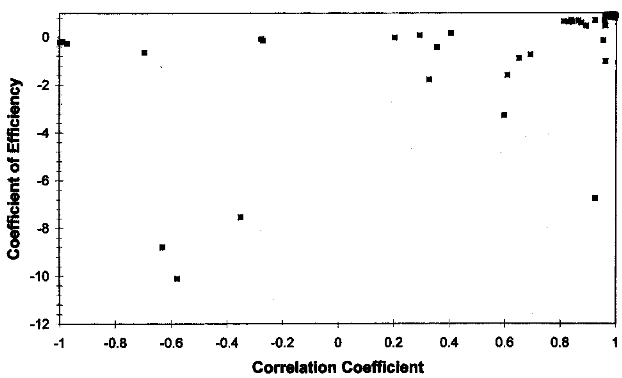

Fig. 1. *Scatter plot between a sample of 50 values of E and R*

adoption of a zero value for E or R as the baseline against which the significance of the goodness of fit is judged. The implication of setting $E = R = 0$ is that an analyst lacking any skill in hydrological modelling would predict a hydrograph which would be a flat, horizontal line drawn through the mean ordinate, i.e. $Q_{est} = Q_{mean}$. For reasons to be explained below, this is an unrealistic null hypothesis which biases the significance of the goodness of fit in the direction of acceptance of a possibly inadequate model, a Type I error.

Improving the baseline

To inject more realism into erecting a baseline, one ought to allow for the fact that the process of transforming effective rainfall into an immediate response runoff hydrograph involves a kernel function with finite area and positive ordinates. This in turn imposes a classic geometrical form on the modelled response of the same shape as is exhibited by actual flood hydrographs, especially those of the simple form customarily used in flood analyses. A more accurate representation of a no-skill scenario ought to allow for this artefact of the modelling process. This can be achieved by measuring predictive improvement over a generalized hydrograph shape with randomly drawn peak and duration.

Sampling distribution of goodness of fit

A simulation exercise was undertaken to replicate the situation where a predicted hydrograph shape is compared with a recorded one. This was achieved by generating pairs of Gamma function-shaped hydrographs, one of each pair representing an observed flood hydrograph, the other representing a model prediction. This entailed random drawings of a standardized width parameter W/Tp and a scale parameter, Tp, where W is the width of the hydrograph containing 50% of its volume and Tp is the time to peak from which Gamma function parameters were computed. W/Tp was taken to be log-normally distributed with parameters chosen to centre values on 2 and to contain almost all values within a range of 0.5 to 8.0. Tp was normally distributed in such a way as to contain values within the range 0.5 to 8 hr with a mean of 4 hr. These values were adopted after consideration of typical cases from the UK Flood Studies Report (NERC, 1975), though do not much affect the conclusions.

The two objective functions, E and R, were calculated from hourly ordinates of each pair of notional observed and estimated hydrographs. Measures taken to avoid inflating these 'fit' coefficients by including long runs of low values in the hydrograph tails were (a) to limit calculation to time steps in which the hydrograph ordinate was not less than 10% of the peak, and (b) consider timesteps up to the point when 90% of the volume had passed. Both observed and estimated hydrographs were renormalised to unit volume in order to simulate the case where the focus is on the flood profile and not on the total volume of runoff. The process was repeated 1000 times in an Excel 7 programme to assemble the frequency diagrams of Fig. 2.

Results of simulation

Figures 2a to 2c summarise the relationship between 1000 pairs of simulated flood hydrographs generated in such a way as to represent a 'skill-free' hydrograph prediction. In this version, 'skill-free' implies no correlation between the parameters used to generate the pairs of hydrographs.

The main point to note is that useful levels of goodness of fit are common even in the absence of any skill in the fitting procedure. This is highlighted by Fig. 2a which shows that about one-quarter of all skill-free predictions throw up a Coefficient of Efficiency of more than 0.8. Similarly one may note from Fig. 2b that almost 60% of predictions will exhibit a correlation better than 0.8. These high values are due to the shared unimodality and tendency to positive skewness of the two curves – both a feature of reality, not an artefact of simulation. Note also the behaviour in Fig. 2c of the Coefficient of Efficiency. There is no lower bound to this measure and negative values are not infrequent – Fig. 2a reveals that over 20% of simulated hydrograph pairs exhibit a value less than –0.5 which is the point where the sum of squares of differences is three times the corrected sum of squares of the hydrograph ordinates.

Correlated-*Tp* scenario

In the version of the no-skill case considered above it is entirely possible that a broad, late-peaking estimated hydrograph be juxtaposed on a rapid response type of recorded hydrograph. This may also be thought too extreme as the location of the rainfall hyetograph along the

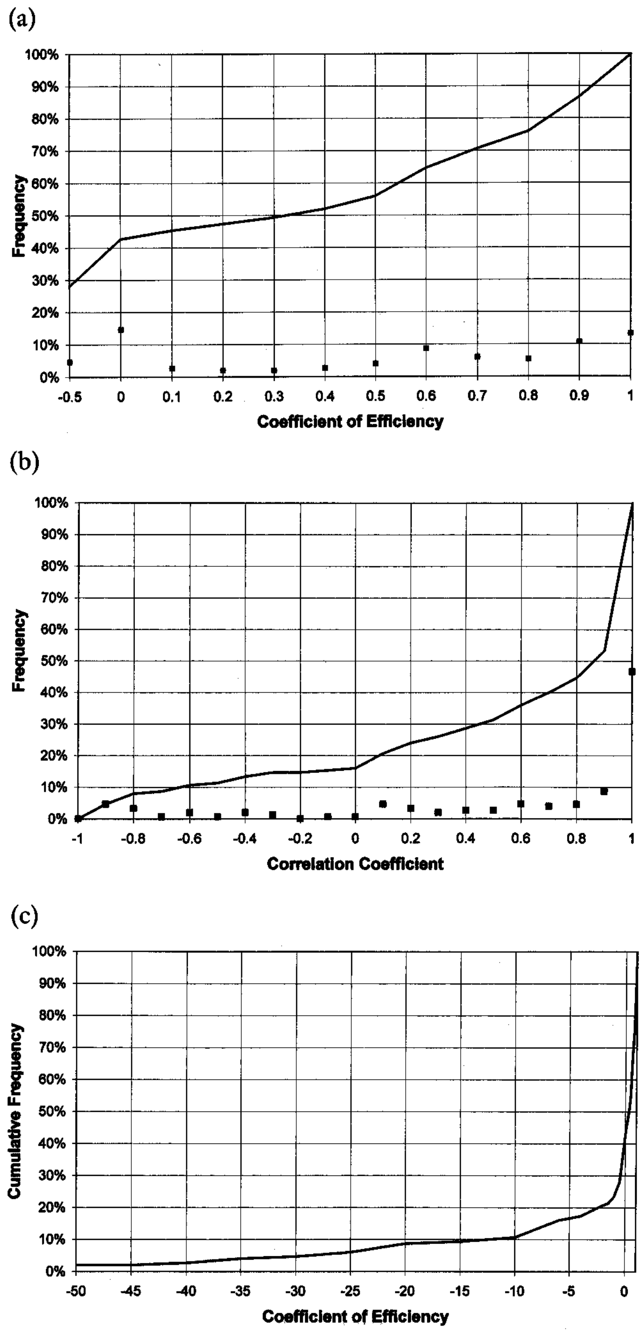

Fig. 2. *Cumulative (line) and relative (points) frequencies curves for a no-skill scenario with uncorrelated Tp: (a) Coefficient of Efficiency; (b) Correlation Coefficient; (c) Coefficient of Efficiency, expanded range from –0.5 to +1.0.*

time axis constrains both observed and predicted hydrographs to produce some degree of alignment of the two. It is a moot point whether this can be described as "skill" as it requires no action on the part of the model-builder in order to incorporate the information. To explore its effect, a second simulation was conducted in which the values of *Tp* of the two hydrographs were correlated, 50% of the variance of *Tp* is explained. A considerable shift in the position of the frequency curves was observed such that two-thirds of correlations exceed 0.8. Correspondingly for the Coefficient of Efficiency measure, almost one-third now exceed 0.8, up from one-quarter in the previous case. A lesser, though still significant number of cases still throw up large negative values of *E*.

Conclusion

This research note has examined the use of goodness of fit measures in the context of prediction of the shape of flood hydrographs. It has been shown that equating the no-skill situation with a horizontal hydrograph through the mean leads to an exaggerated impression of the presumed skill in prediction. Unfortunately this is the null-hypothesis that is implied in past studies. The alternative scenarios explored here have allowed for the built-in tendency for model output and recorded hydrographs to share basic geometrical features. Figures presented here indicate that, to be confident that the model is genuinely adding value, one should not be satisfied with a Coefficient of Efficiency lower than mid to high 0.9s.

References

Binley, A.M., Beven, K.J., Calver, A. and Watts, L.G., 1991. Changing responses in hydrology: assessing the uncertainty in physically based model prediction. *Wat. Resour. Res.* **27** 1253–1261.

Gupta, H.V., Soroooshian, S. and Yapo, P.O., 1998. Towards improved calibration of hydrologic models: multiple and non-commensurable measures of information. *Wat. Resour. Res.* **34** 751–763.

Jakeman, A.J. and Hornberger, G.M. 1993. How much complexity is warranted in a rainfall-runoff model? *Wat. Resour. Res.* **29** 2637–2649.

Nash, J E and Sutcliffe, J.V. (1970) River flow forecasting through conceptual models. Part 1 – A discussion of principles. *J. Hydrol.* **10** 282–290.

Natural Environment Research Council. 1975. Flood Studies Report. London.